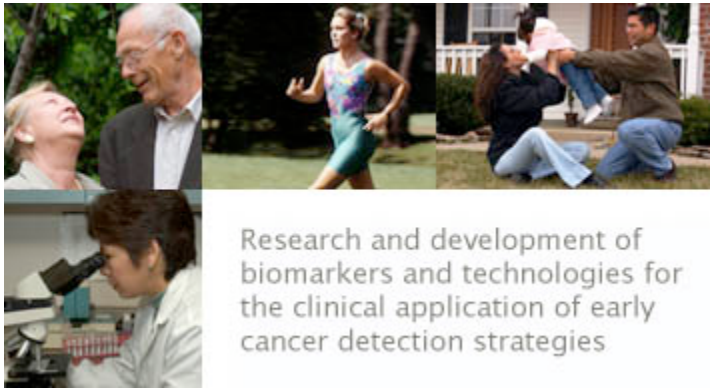


FRED HUTCHINSON CANCER RESEARCH CENTER
A LIFE OF SCIENCE

Early Detection Research Network

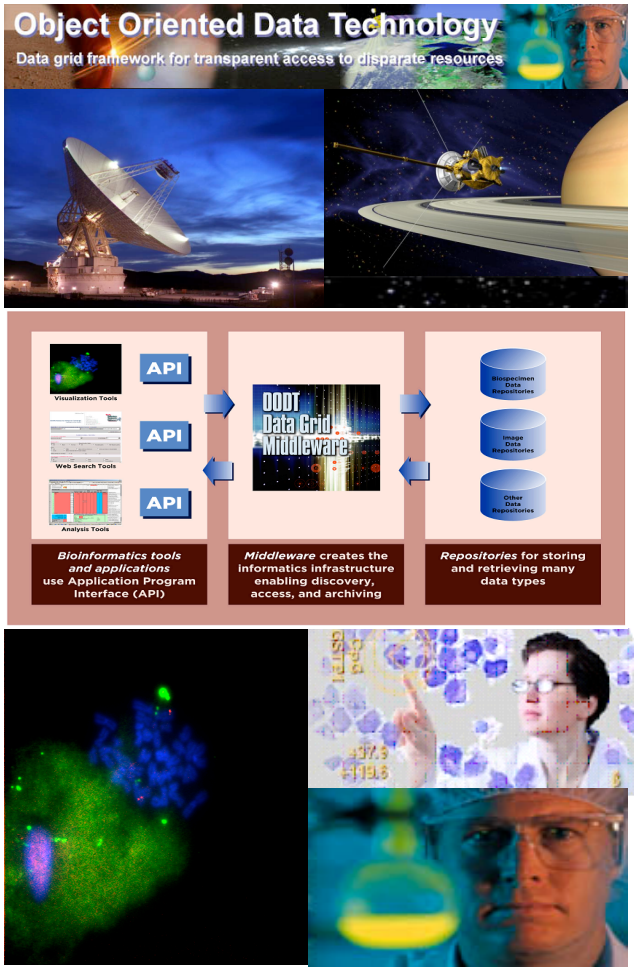
JPL
Jet Propulsion Laboratory
California Institute
of Technology

NCI EDRN Biomarkers Knowledge Environment Informatics Infrastructure

Sudhir Srivastava
Chief, Cancer Biomarkers Research Group
Division of Cancer Prevention
National Cancer Institute

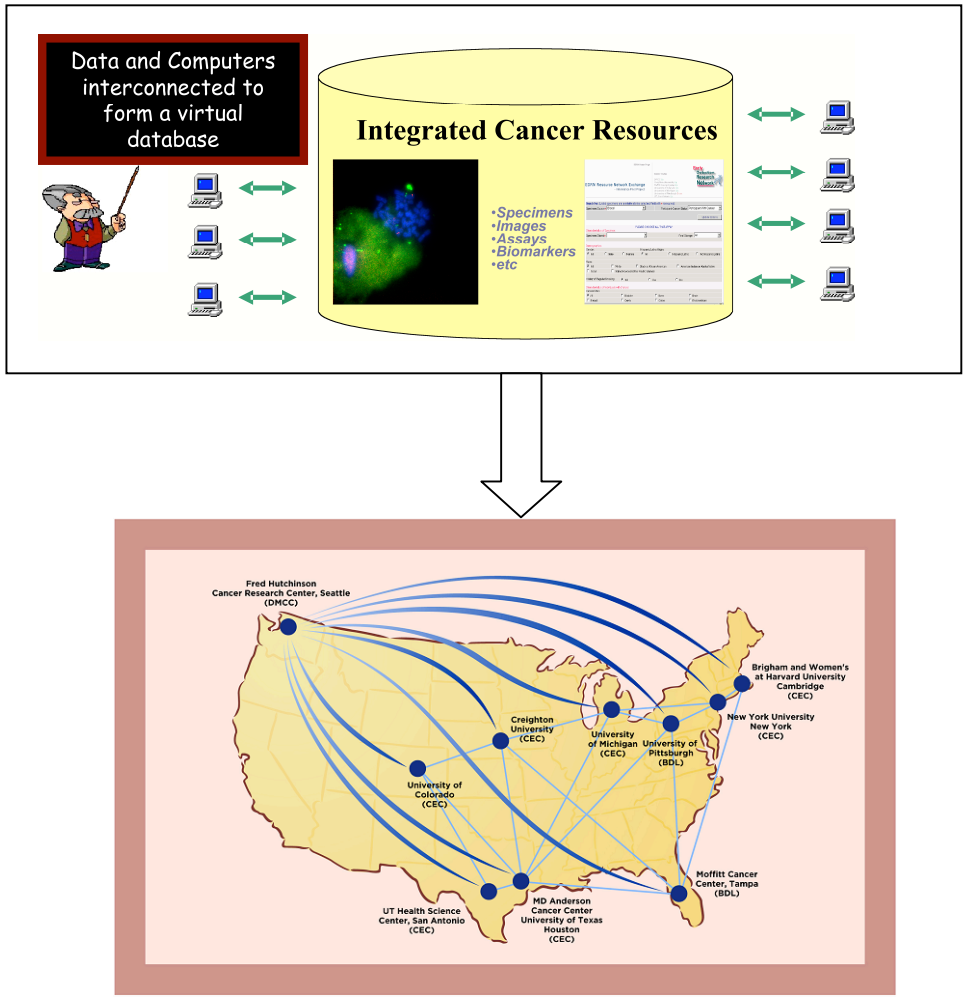
Dan Crichton
Program Manager for Engineering, NASA Planetary Data System
PI, Informatics, NCI EDRN Program
Jet Propulsion Laboratory

July 28, 2007



The Informatics Challenge…

- Distributed scientists across ~40 research centers
  - Parallel and on-going research
  - Diverse data sets in highly distributed catalogs
- Science discovery through capture, mining and correlation of diverse data sets acquired during EDRN validation studies
- Sharing of data resources between diverse, distributed science research databases
  - Biomarkers
  - Proteomics
  - Biospecimens
  - Various technologies and data products (image, micro-satellite, …)



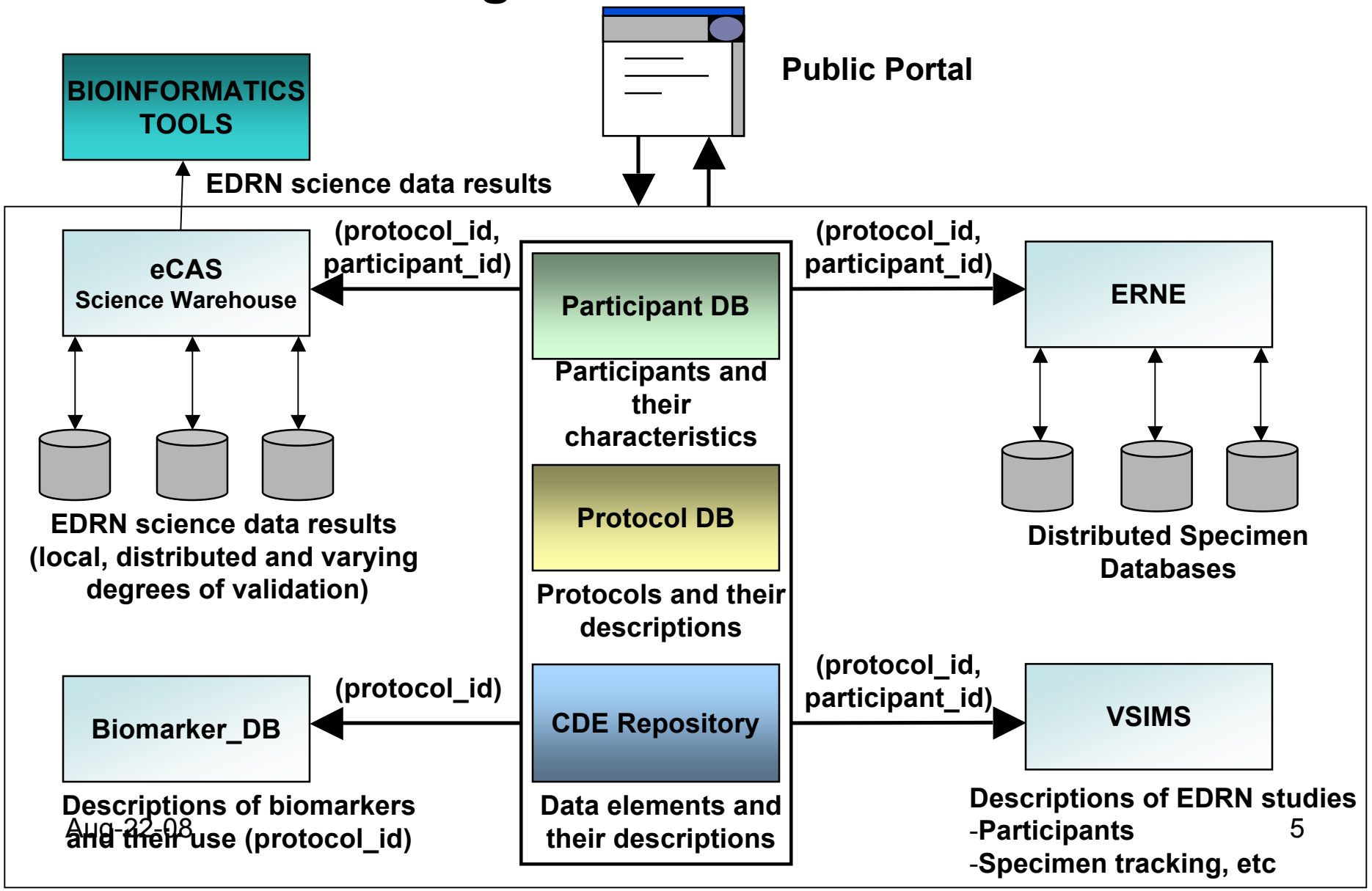
Original Informatics Vision for EDRN

- Advances in distributed computing infrastructures provide
  - Connectivity of computers and repositories globally over the Internet
  - Electronic sharing of data
  - Web-based access to data distributed in independent databases
  - Software and data standards for the access and sharing of data
  - Global security mechanisms to protect data sharing and access
- In effect, EDRN was pursuing a new paradigm for biomedical research in which data and computing can remain distributed, but be integrated into a virtual knowledge environment

EDRN Informatics Goals and Principles

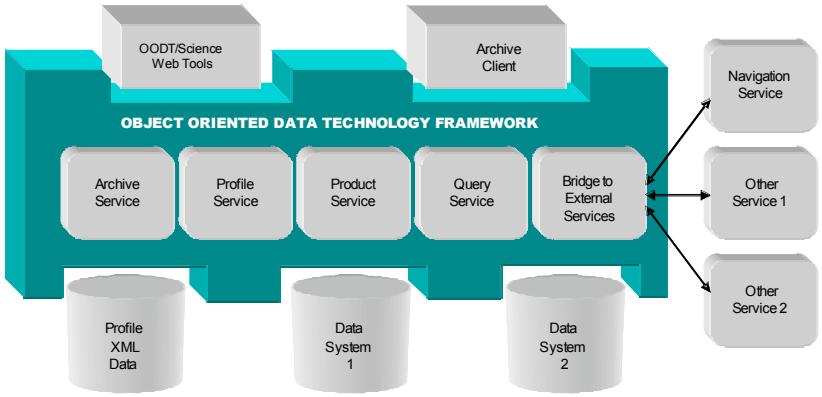
- Develop a knowledge system that **links** together EDRN data assets into a virtual data system based on **common data elements**
- Establish an EDRN bioinformatics program that promotes the use of a **common informatics infrastructure** by EDRN sites.
- Provide an infrastructure for **capturing** EDRN validation study results and a mechanism for **distribution**
- Define data and software **standards** for EDRN informatics systems
- **Collaborate** with both EDRN and non-EDRN sites on informatics.
- Develop a public portal that provides **information dissemination** about EDRN programs and progress.
- Enable tools that support **scientific inquiry** both within and across databases and data sets.

Realizing the Informatics Vision



Leveraging Software from NASA: Object Oriented Data Technology

- Started in 1998 as a research and development task funded at JPL by the Office of Space Science to address
  - Application of Information Technology to Space Science
  - Provide an infrastructure for distributed data management
  - Development of a set of software components to support generation and sharing of data for distributed science domains



- OODT Initial focus on capturing and sharing data across distributed science data repositories and systems
- Runner-up NASA Software of the Year, 2003
- Used in planetary, astrophysics earth and biomedical sciences
- Java framework available via Open source at http://www.openchannelsoftware.com

Sharing Specimens: A start at building the infrastructure

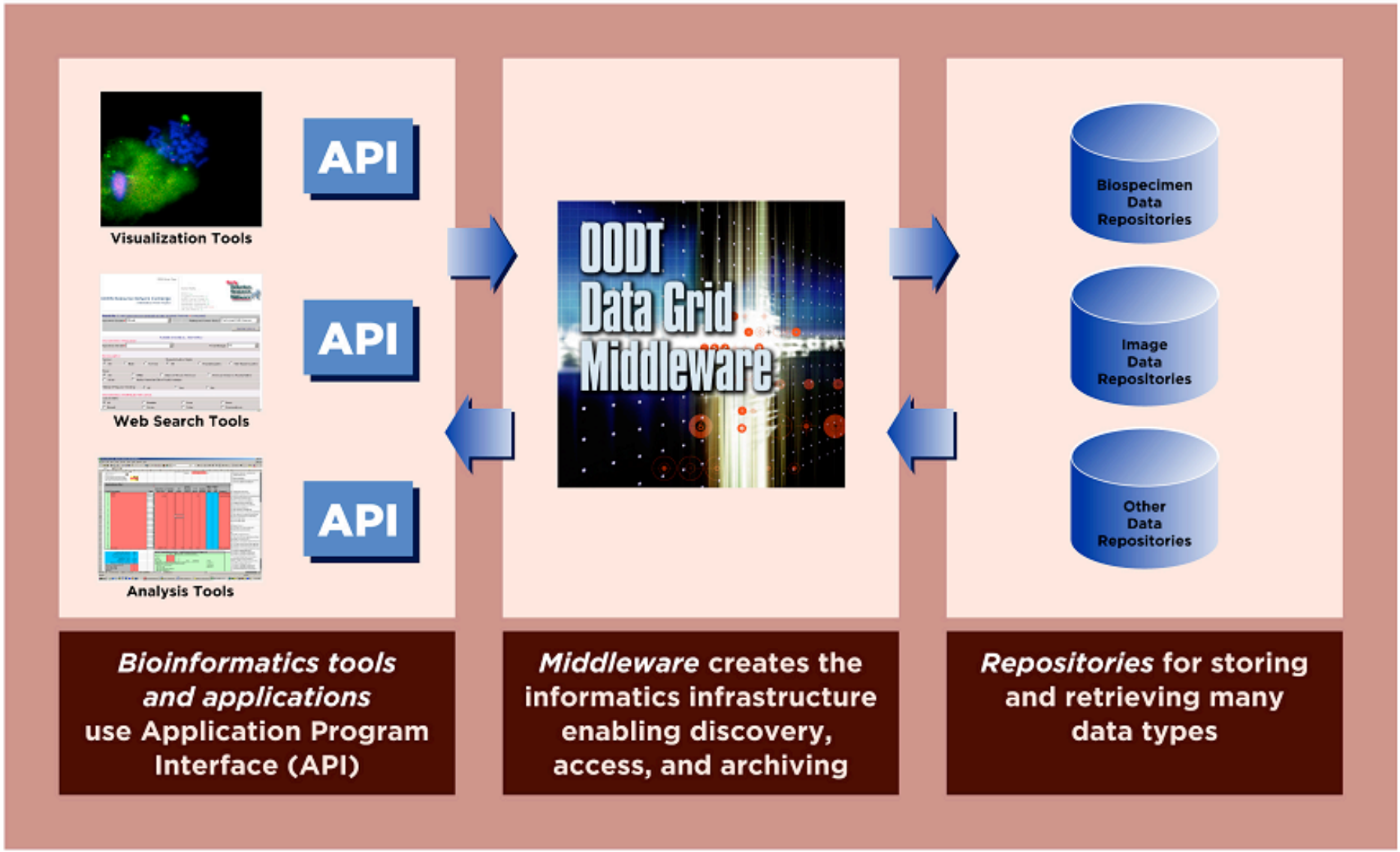
- Specimen management is a critical function that is carried out across biomedical research centers
  - Capturing the characteristics of the specimens (e.g., their epidemiological characteristics) is critical
  - Scientists need access to certain specimens during an assay or study
- Access to a "virtual specimen bank" viewed as important in supporting EDRN's need for collaboration within the network
- But, challenges remained…
  - No standards for how specimen information is represented…often homegrown
  - Very different levels of technology, support, etc at sites
  - Specimen management is reviewed and controlled by Institutional Review Boards (IRBs) put into place by the federal government

Project 1: EDRN Resource Network Exchange (ERNE)

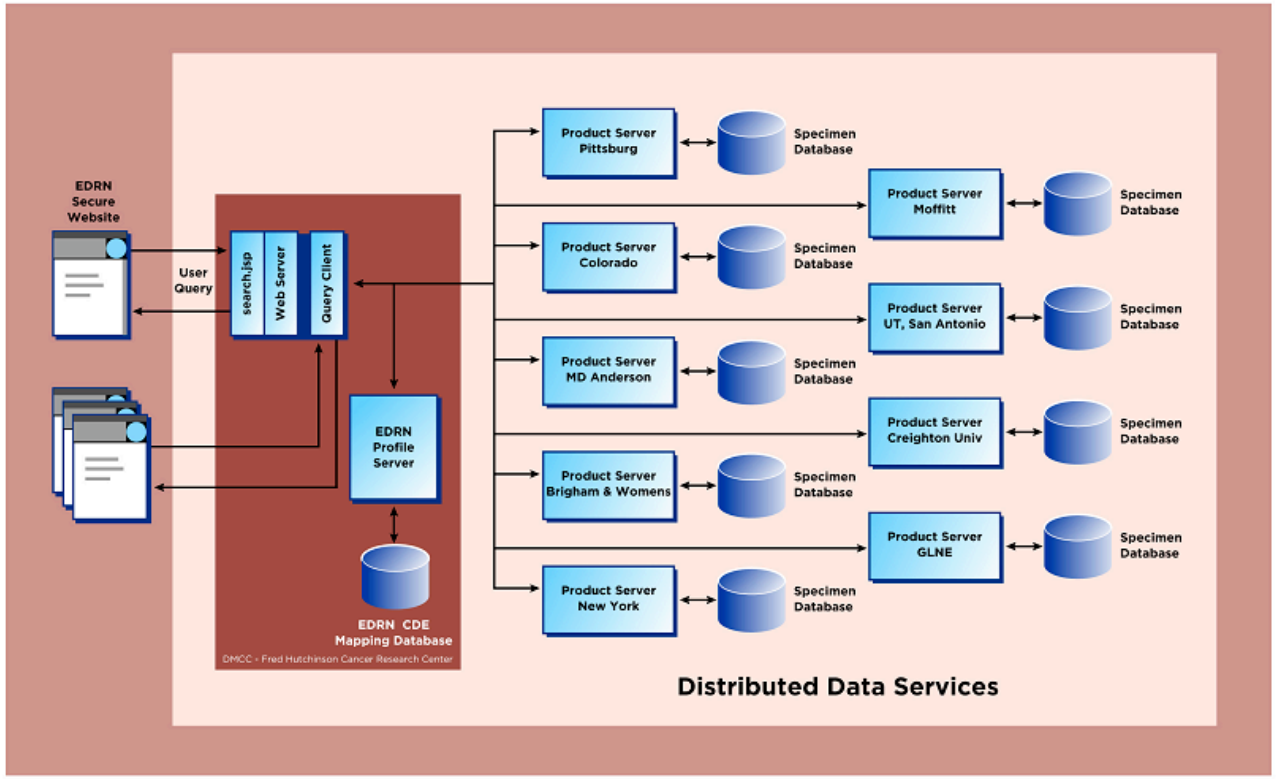
- Initiated as a proof of concept to construct a virtual specimen system linking databases together from major cancer centers
  - Linked over 10 sites (on our way to 15)
  - Developed in three phases starting with just two sites
- Established a cross-disciplinary team including
  - Scientists (from cancer centers)
  - Policy Makers (from NCI and NIH)
  - Coordinators (from Fred Hutchinson Cancer Center in Seattle/EDRN Data Management and Coordinating Center)
  - Technologists (from JPL)
- Became an NCI Success Story…
  - 2002 Article in Journal of National Cancer Institute
  - Initiated National Biospecimen Network (NBN)

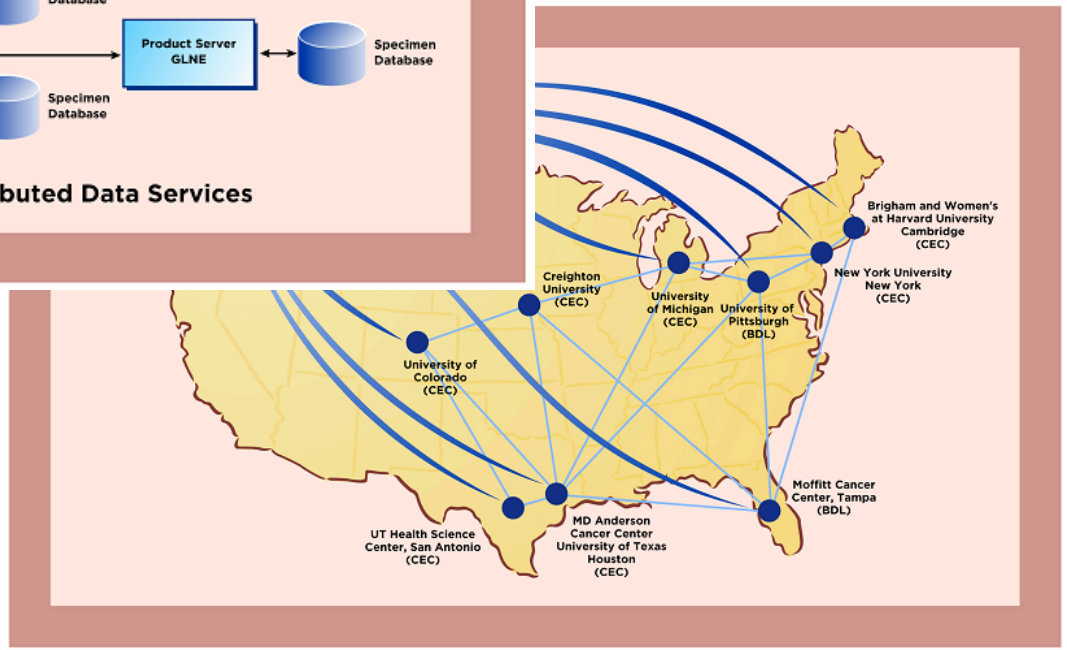
Early Detection Research Network

Distributed Access Model

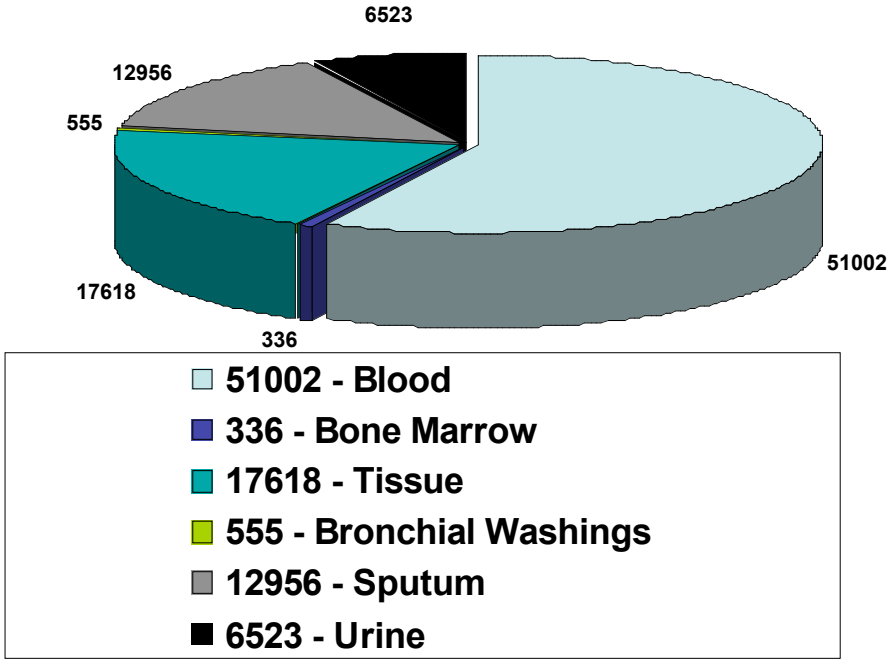


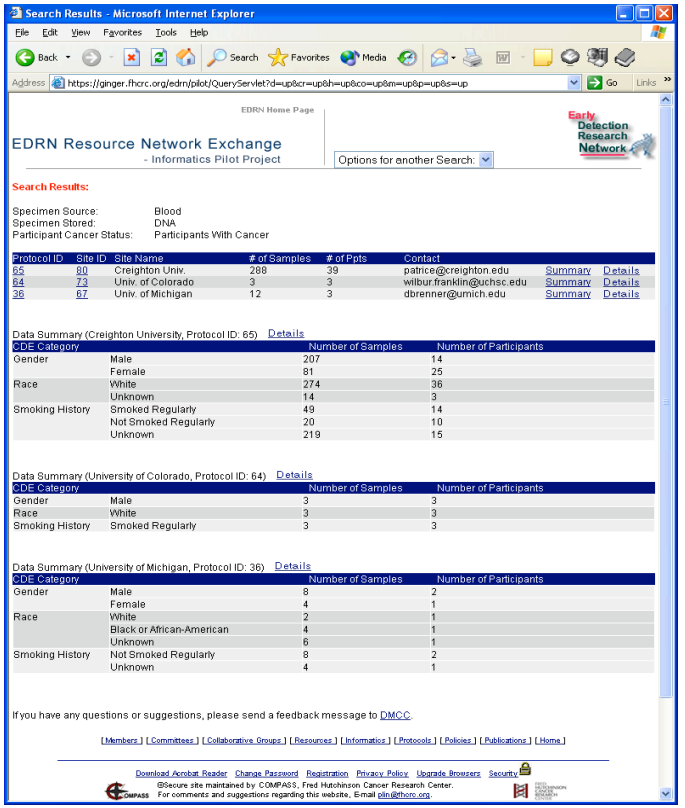
EDRN Specimen Integration and Deployment





Types of Specimens in ERNE

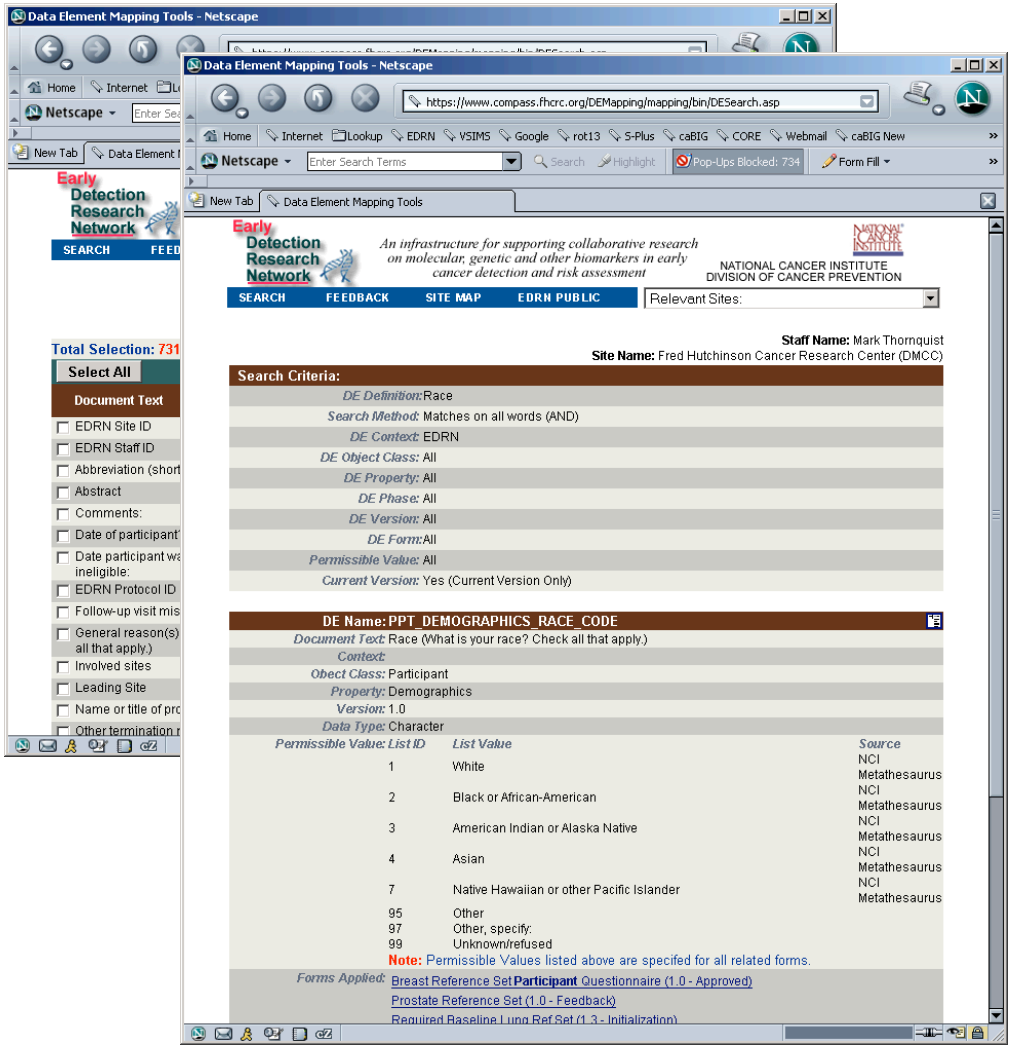




Project 2: Common Data Elements

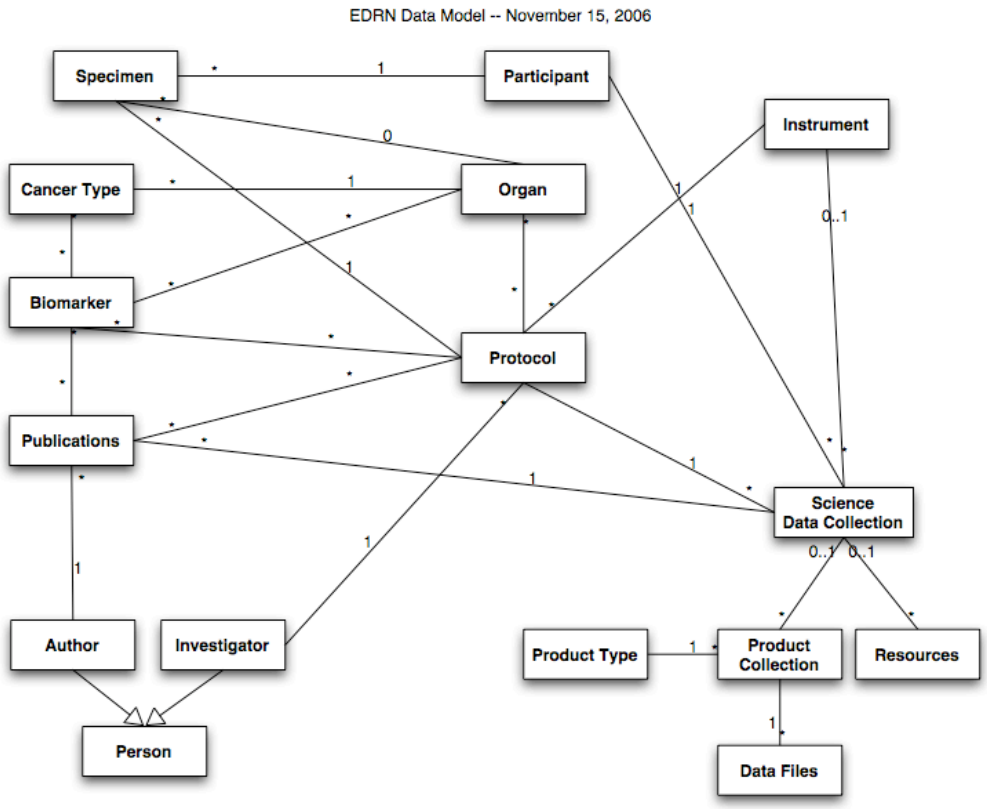
- Common Data Elements provide a set of standard terms and values for a domain
  - They are classified into organ, epidemiological and specimen CDEs
  - **Critical to getting ERNE to work**, but have historically been forms-based (lack an overarching information model)
  - Based on ISO/IEC 11179 (standard for data elements)

- Captured by EDRN and maintained by the EDRN Data Management and Coordating Center in Seattle

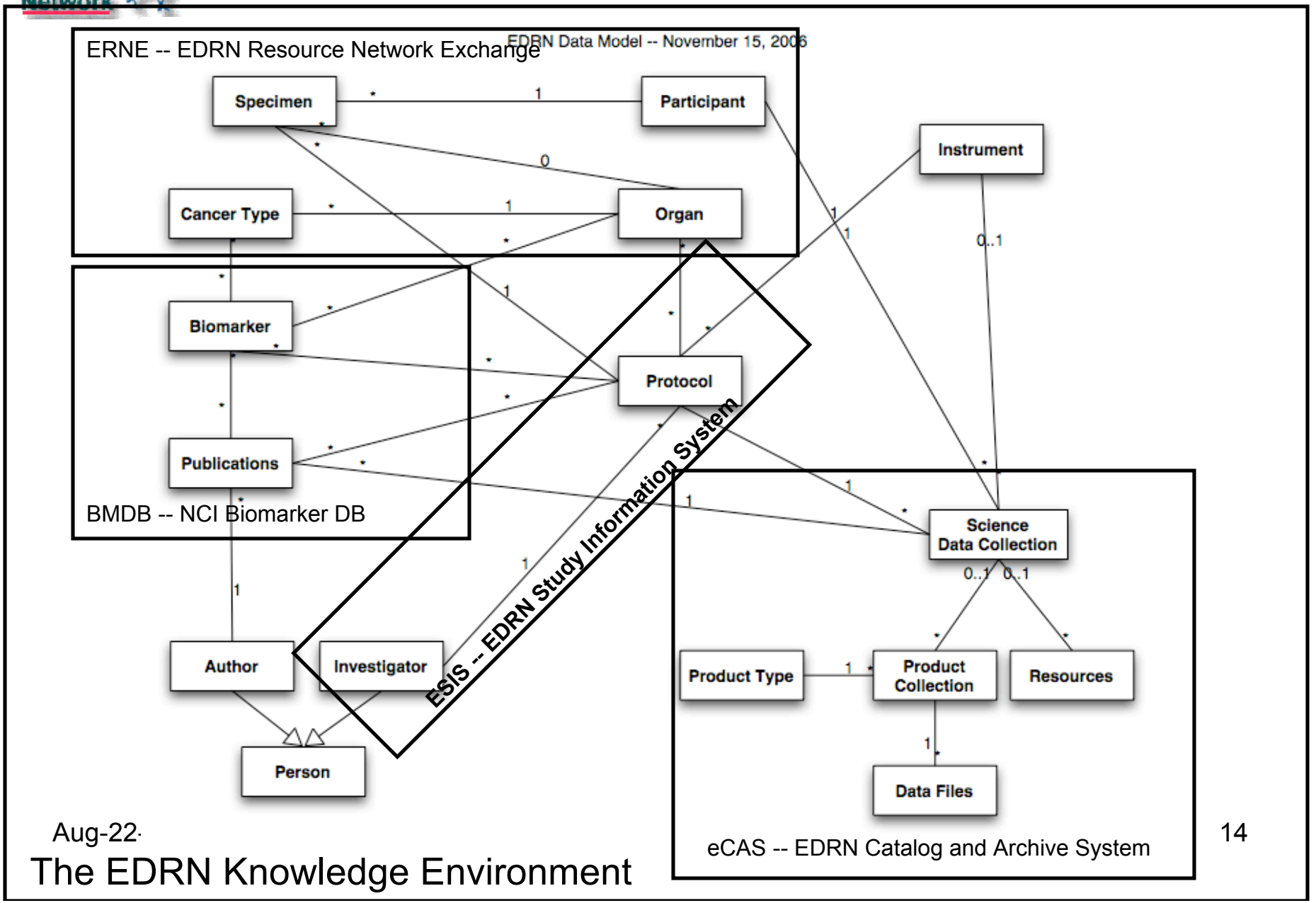


Project 3: EDRN Information Model

- High level *ontology* model of EDRN which describes
  - Core "object" concepts of EDRN data
  - Relationships between those objects
- Specific models are derived from this high level model
  - Model of biospecimens
  - Model for each class of science data
- EDRN is specifically focusing on a granular model for annotating biomarkers and their studies
- But, the high level model allows us to move towards an integrated information environment
- **<u>The model is independent of the software component architecture</u>**



EDRN Data Model Mapping to Applications

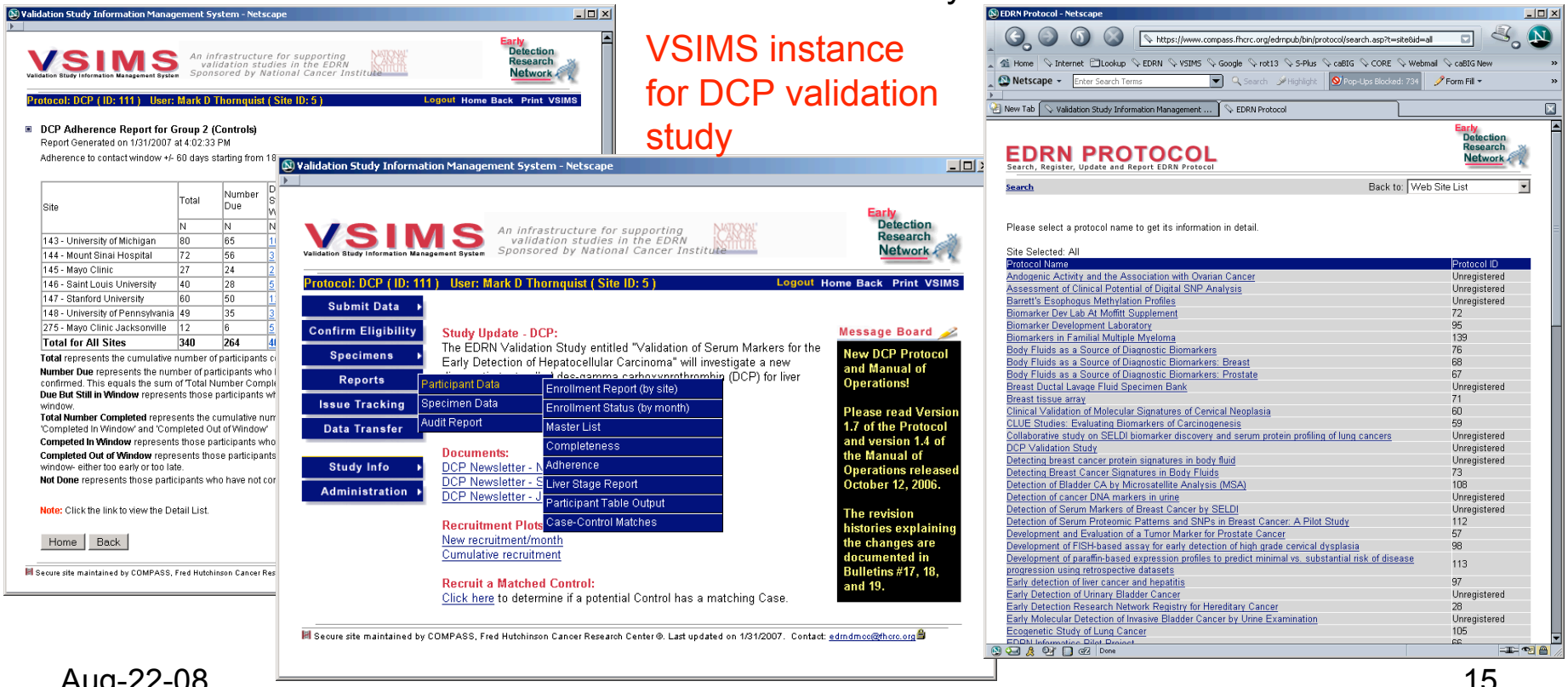


Project 4: Validation Study Management

- **VSIMS:** Online study management system supporting all EDRN validation studies. Built upon the EDRN CDE repository and using reusable modules to speed development for new studies.

eSIS: System in development to track the progress of all EDRN-funded projects, including timelines, GANTT charts, phases of development, current study status

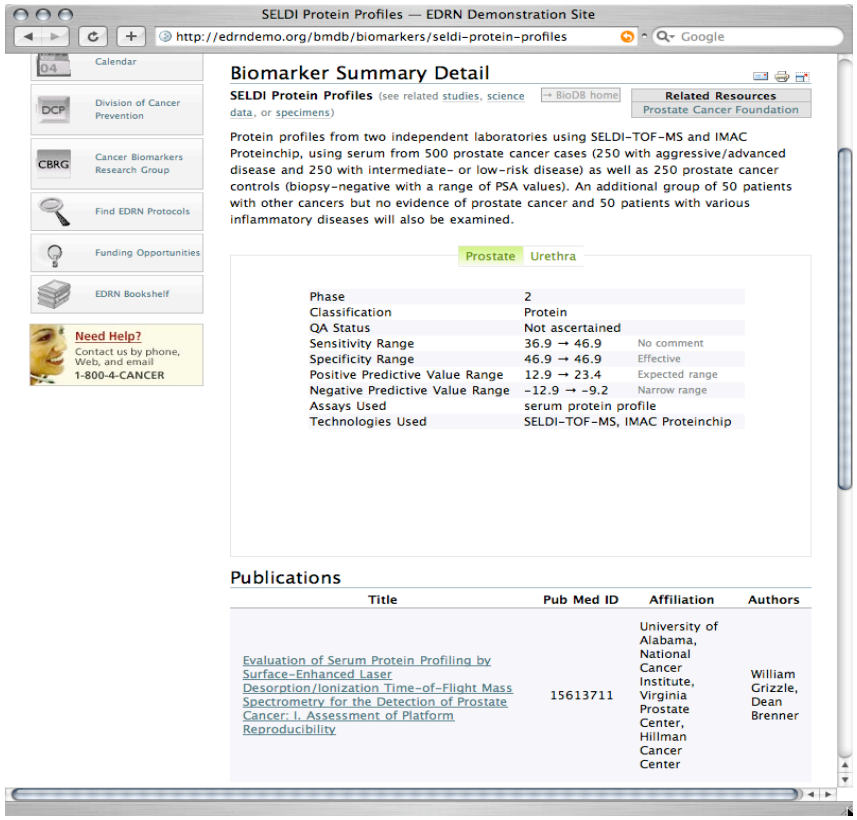
VSIMS instance for DCP validation study



Project 5: Biomarker Database

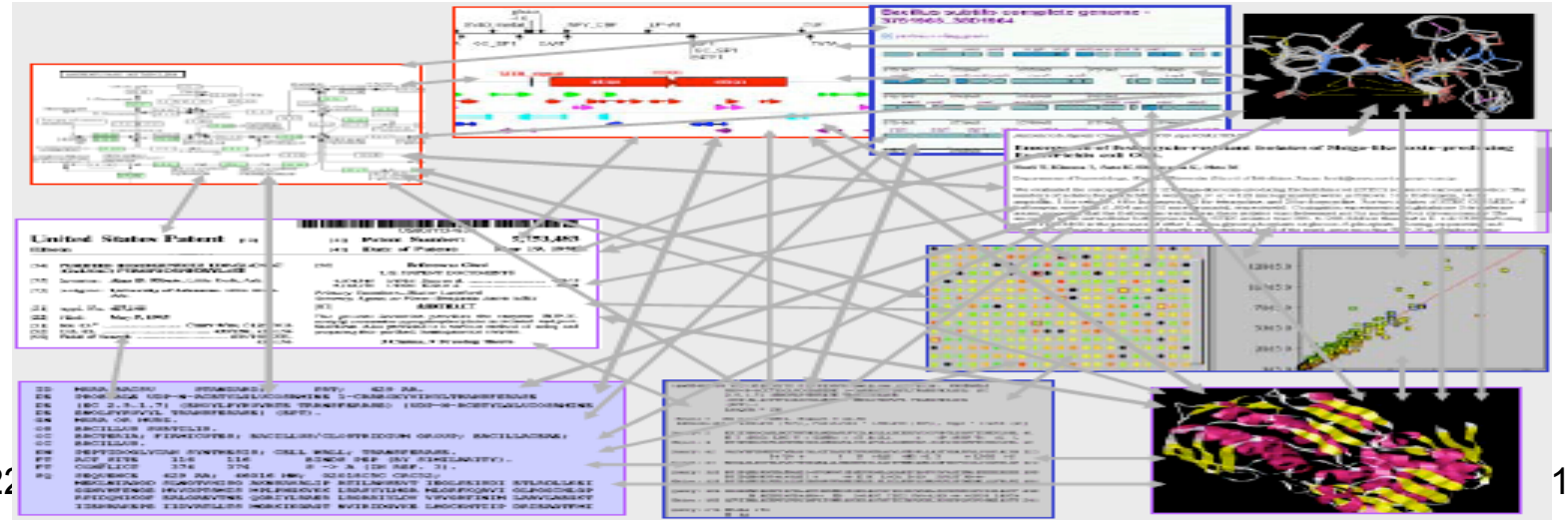
- Develop a registry to annotate biomarkers that are either under development or reported in publications
  - Entry into the registry would be through peer review
  - Initiated by EDRN, but open to NCI

Biomarker Descriptions → Biomarker Studies



Biomarker Database Capabilities

- Tracking of Biomarker Research Progress
  - Report research progress through phases of early detection research
  - Monitor multi-function, multi-institution researches
- Integration of Biomarker Information
  - Provide a means for effective and easy utilization of data, communicating new discoveries
  - Provide links to the relevant literature reference and the appropriate genomic and proteomic databases for the relevant information for each biomarker



Current Status of Biomarker DB

- V1.0 of the model is complete
  - Capture of information is now underway
- A pilot database has been developed and integrated into the EDRN science portal (more on that later…)
  - Will provide a beta test release in September as part of the EDRN Public Portal
- EDRN is establishing assembling a curation process and group for the database

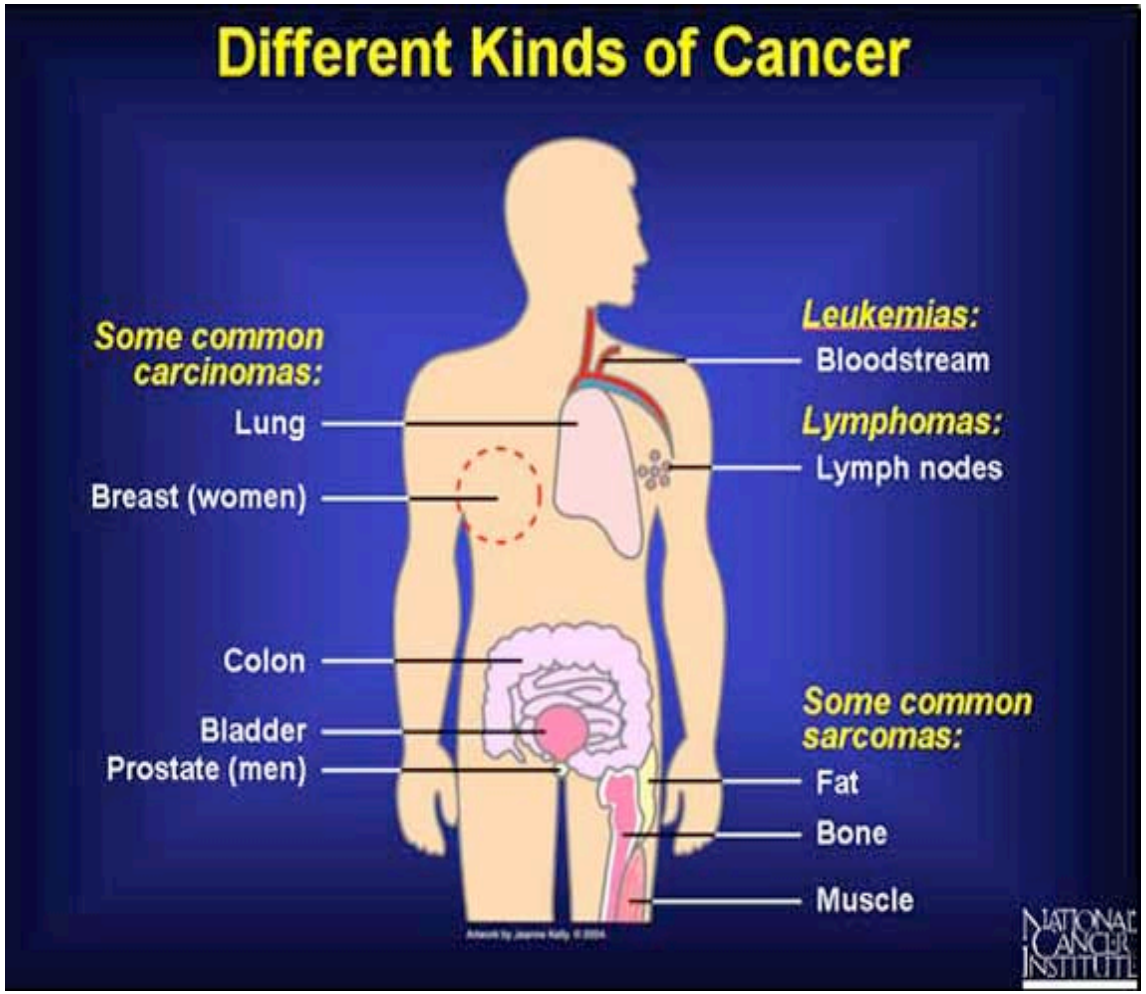
Project 6: Science Data Warehousing and Access

- Provide an integrated repository of EDRN data sets (e.g., mass spectrometry data, microarray data, 2-D electrophoresis gels, etc.) released by investigators
  - Use the EDRN CDEs to populate a catalog describing the data sets
- Provide tool set for constructing (managing and generating) biorepositories
- Provide a distribution mechanism to the community for EDRN public science data
  - Provide granular searches across distributed data sets, integrated into the EDRN knowledge system
- Provide long term preservation of EDRN study information

EDRN Science Data Covers Different Types of Cancer

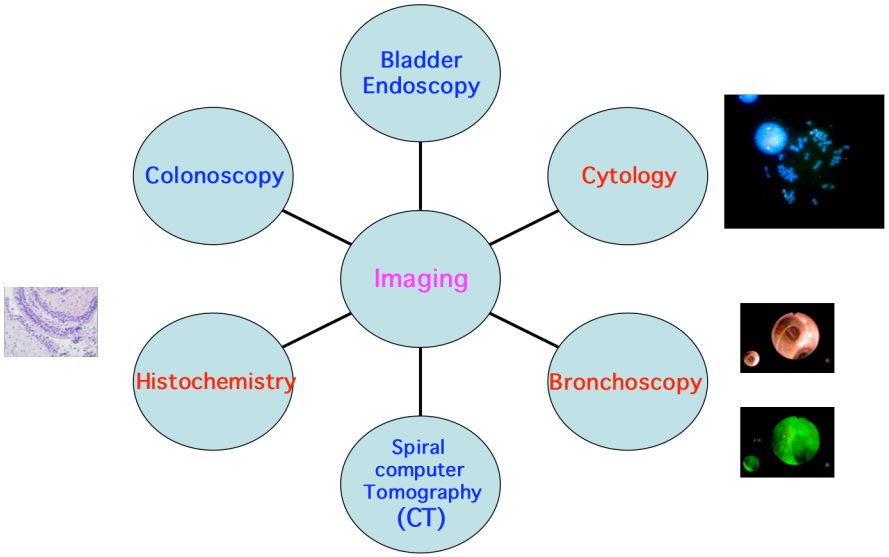
Common Cancer Types

- Bladder Cancer
- Breast Cancer
- Colon Cancer
- Endometrial Cancer
- Kidney (Renal Cell) Cancer
- Leukemia
- Lung Cancer
- Melanoma
- Non-Hodgkin's Lymphoma
- Pancreatic Cancer
- Prostate Cancer
- Skin Cancer (Non-melanoma)
- Thyroid Cancer
- Ovarian Cancer
- Liver Cancer

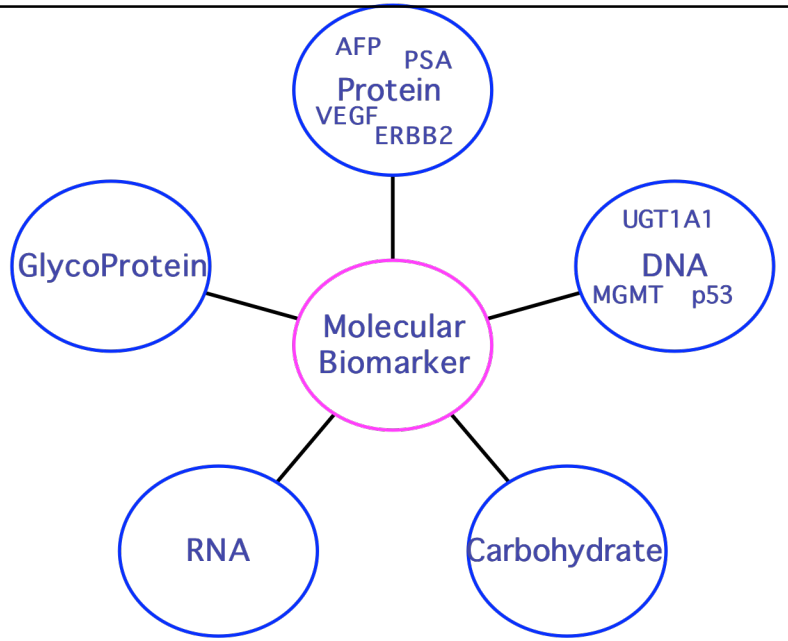


Managing EDRN's Multi-disciplinary Science Data

EDRN's underlying information model needs to be flexible in order to manage a variety of different types of science data captured by different experiments at different phases in studying biomarkers…

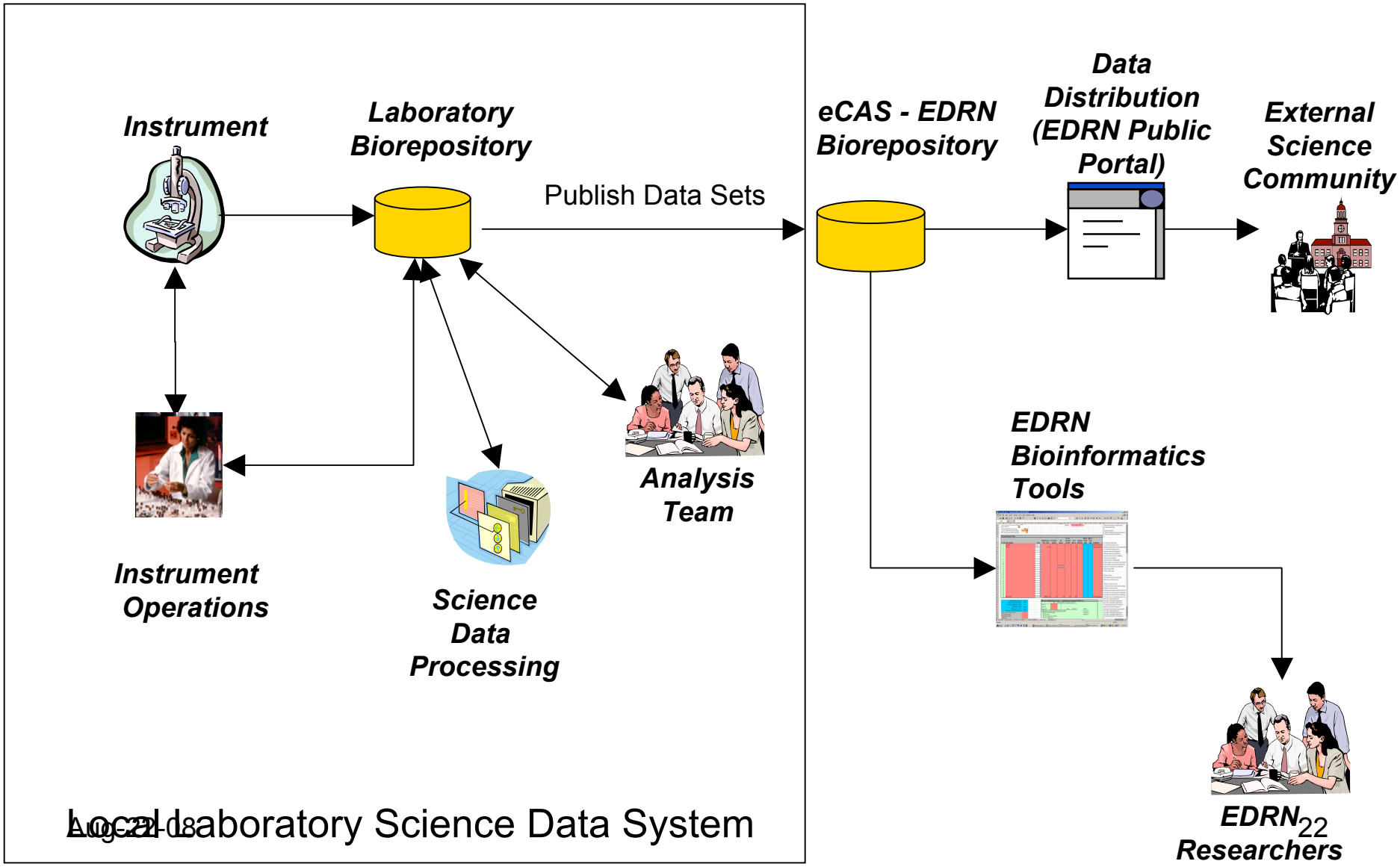


Imaging Data Types



Different Data Types

EDRN Science Pipeline



The EDRN "Biomarker Atlas"

- Focus on organ centric discovery and access to science data
- Distributed Biomarker Atlas for the Lung
  - Prototype with access to distributed lung image databases (Colorado and Roswell Park)
  - Registration of science products against lung map
  - Integration of distributed image and specimen data sets
  - Demonstrated at the Lung SPORE meeting in July by Wilbur Franklin

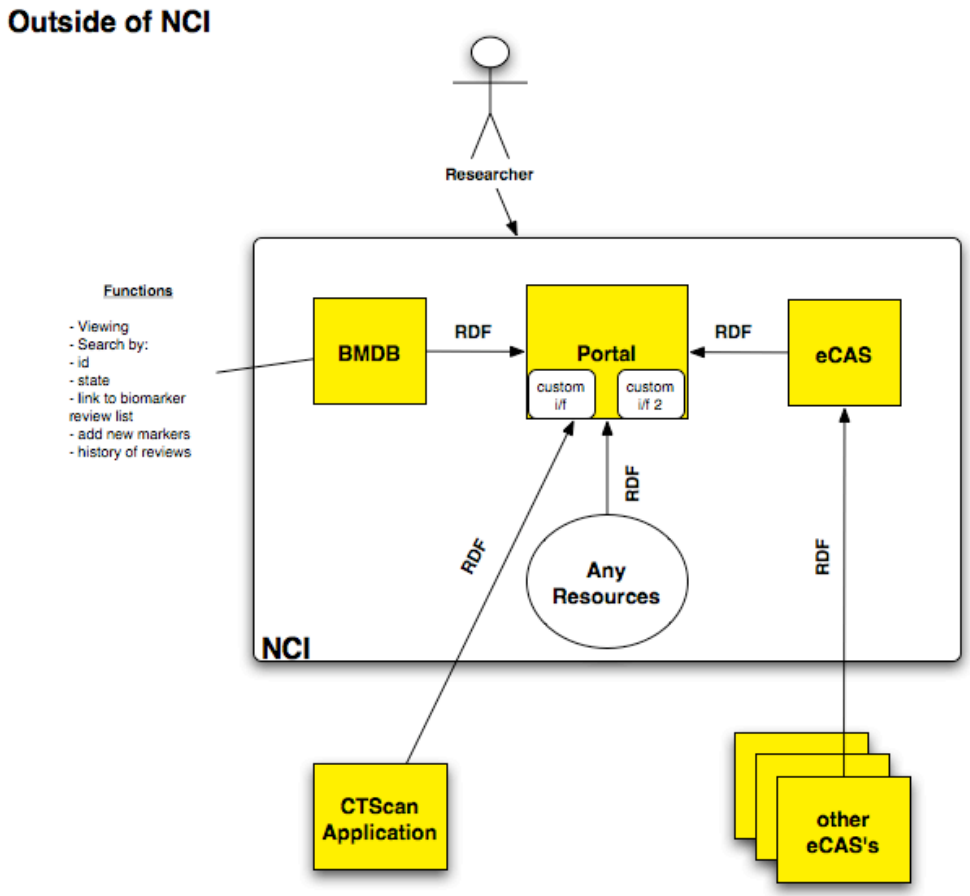
Biomarker Atlas for Lung

Project 7: Public Portal

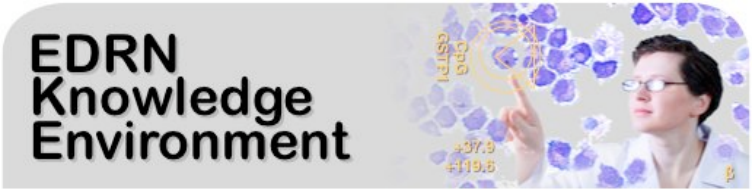
- Originally released in 2005
  - Built on Plone - open source portal software
- Used by NCI and EDRN for disseminating program information
- But, will migrate to a science portal soon…
  - Currently under development

Moving to an Integrated Science Architecture

- Semantic science portal driven by the EDRN ontology
  - Schema loaded into the ontology via RDFS (and Protégé)
  - Metadata from distributed applications dumped into the portal via RDF



Integrated EDRN Knowledge Environment



- Model-driven Portal
  - Adapts to a dynamic set of “object types” (mass spec, immunohistochemistry, etc) based on the model
  - Provides access to distributed repositories of information
  - Provides “google-like” search of the object types

Progress in cancer research towards science-driven informatics architectures

- Recognition of how to architect science-driven distributed software systems*
  - Separate the architecture into core pieces (process, data and software)
  - The “information model” is critical
    - Should provide a generalized mechanism to describe and organize data
    - Model-driven systems provide the agility to support multi-project, multi-center studies
  - Develop modular software components that can be configured based on the “information model”
    - Modularity helps to drive both longevity and agility in system designs
  - Allow for geographically distributed software components to communicate based on standards
  - Identify and implement core scientific “use cases” that help to evolve the system
  - EDRN has demonstrated this architecture can work in managing and sharing specimen information
  - JPL has done this for planetary science and is now working with international space agencies to provide access to scientific data results returned from international missions

*D. Crichton, S. Kelly, C. Mattmann, Q. Xiao, J. S. Hughes, J. Oh, M. Thornquist, D. Johnsey, S. Srivastava, L. Esserman, W. Bigbee. **A Distributed Information Services Architecture to Support Biomarker Discovery in Early Detection of Cancer.** In Proceedings of the *2nd IEEE International Conference on e-Science and Grid Computing, pp. 44*, Amsterdam, the Netherlands, December 4th- 6th, 2006.

More Lessons Learned…

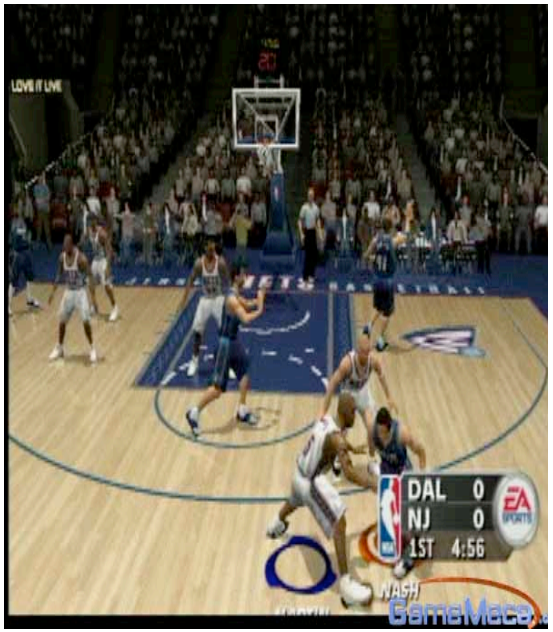
- Technology needs to be accessible (with a different levels of entry)
  - We need to work with sites to help them understand technology and use it compliant with federal government regulations
  - We can't do technology for technology-sake. We need to "enable" science and demonstrate its value.

- NASA and NIH science challenges are similar
  - The science is always evolving
  - Distributed, PI-driven, complex data structures, computationally-intensive, etc
  - Need for software that supports automated pipelines
  - The need to capture and share data within scientific communities

- ***Success has been the result of having a "focused approach" to informatics and building cross-disciplinary teams…***

EDRN Informatics can pull it together…

"The most successful and efficient research about molecular markers will require effective interdisciplinary communication and collaboration involving fields of molecular biology, observational epidemiology and biostatistics."



Ransohoff, Nature Rev Cancer 2004; 4:309-314

Acknowledgements

- Mark Thornquist and members of Data Management and Coordinator Center at the Fred Hutchinson Cancer Research Center
- Sudhir Srivastava and Don Johnsey, National Cancer Institute
- The EDRN informatics advisory group Bill Bigbee, Laura Essermann, Wilbur Franklin, Tony Hollingsworth, Jeffrey Marks
- Currently integrated sites:
  - H. Lee Moffitt Cancer Center
  - University of Texas, San Antonio
  - Creighton University
  - University of Colorado
  - University of Pittsburgh
  - University of Michigan/Dartmouth University (Great Lakes New England Consortium)
  - Brigham and Womens
  - MD Anderson
  - New York University
- NASA Jet Propulsion Laboratory